Goal sequence RL
Robot arm pick-and-place

By Phuc Nguyen

Motivation

- Automation and Robot become more and more popular
- Reinforcement learning is the main momentum of Robot control
- A simple solution with encapsulated production-ready modules

CNBC

History

1. Standard two-step approach: object recognition and pose estimation followed by model-based grasp planning
2. HER can resolve sparse reward problem; shaped reward function might be detrimental to the algorithm and requires lots effort
3. Offline RL, the accuracy of the value estimates depends on the richness of the dataset in terms of its state and action space coverage

Some related work in the area

1. Data-driven Deep semantic segmentation (Wong et al.)
    a. ConvNets to provide bounding box proposals or segmentations, followed by geometric registration to estimate object poses
2. Transporter network (Zeng et al.)
    a. A simple model which can rearrange deep features to infer spatial displacements from visual input, which can parameterize robot actions

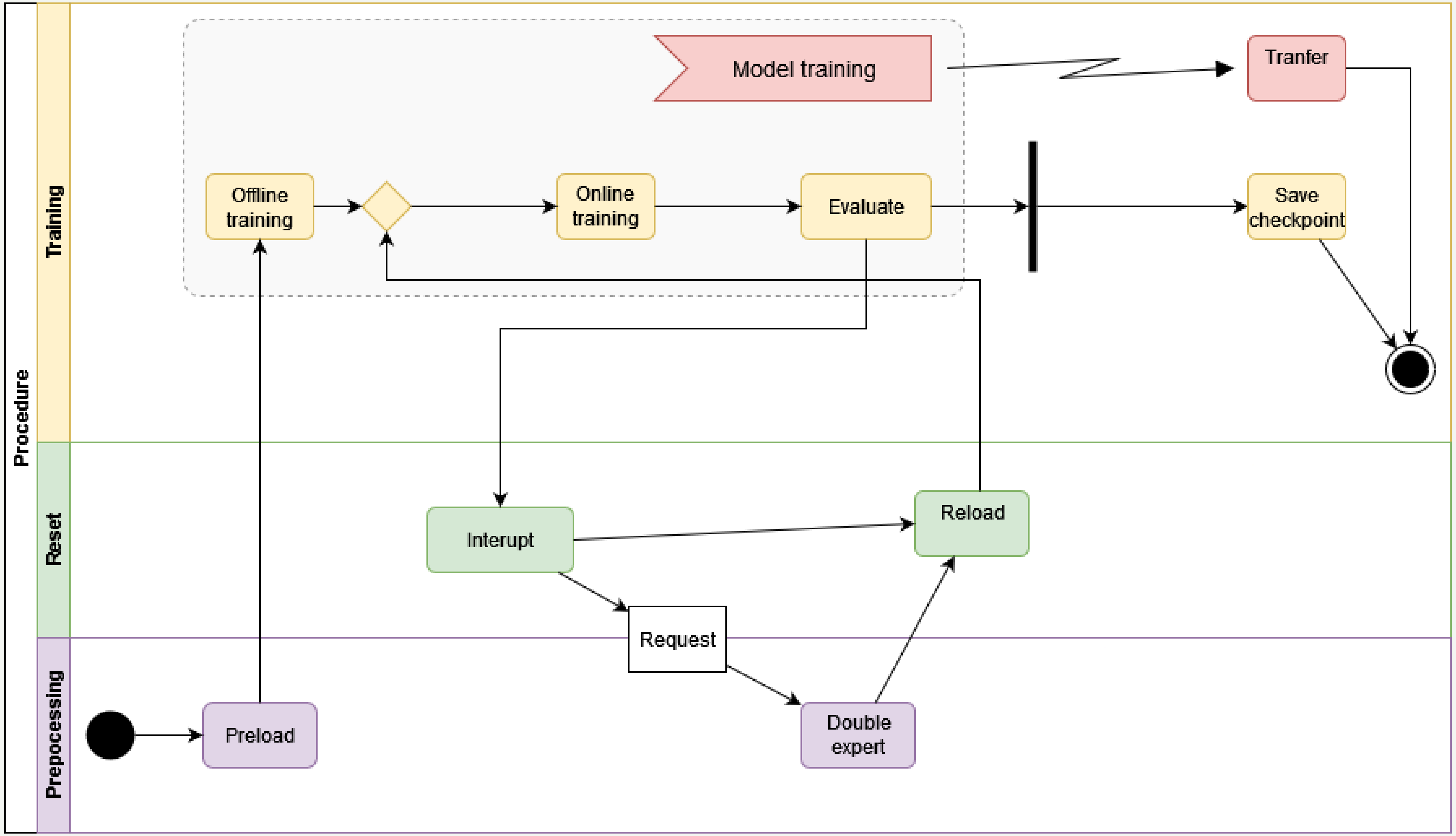
Slide on High level idea for method

- Stable-baseline Truncated Quantile Critics
  - Production ready, huge community support
  - Replay buffer with offline learning capacity
- Offline learning and gradient step
  - Provide warm start
- Transfer learning and domain randomization
  - Support Sim2Real and multi-object

Procedure

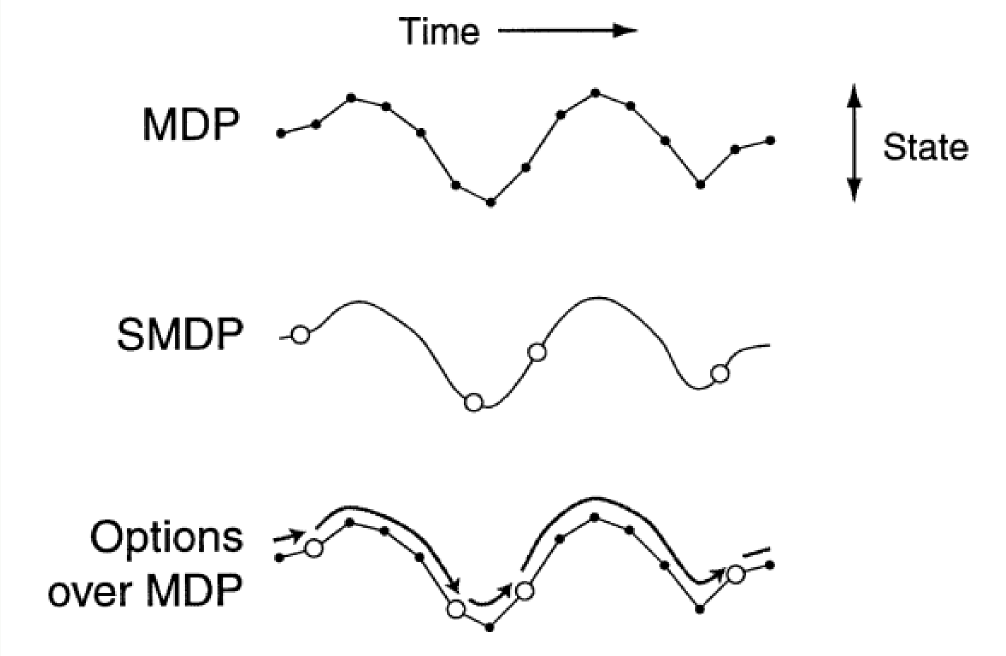
Guided learning

Semi-online RL



Hierarchical Reinforcement Learning

1. Avoid goal-conditioned RL due to complexity in deployment
2. Options framework to get better generalization
3. Guided online PPO
   1. Alternate training procedure
   2. Enforce option complexity
   3. Dealing with gradient explosion



Doina Precup

Experiment

1. Single object
    a. Guided learning and advanced HER
2. Multi-object
    a. Transfer learning
    b. Inverse RL/shaped reward engineer
    c. Massive expert data

Algorithm 3 Advanced HER algorithm

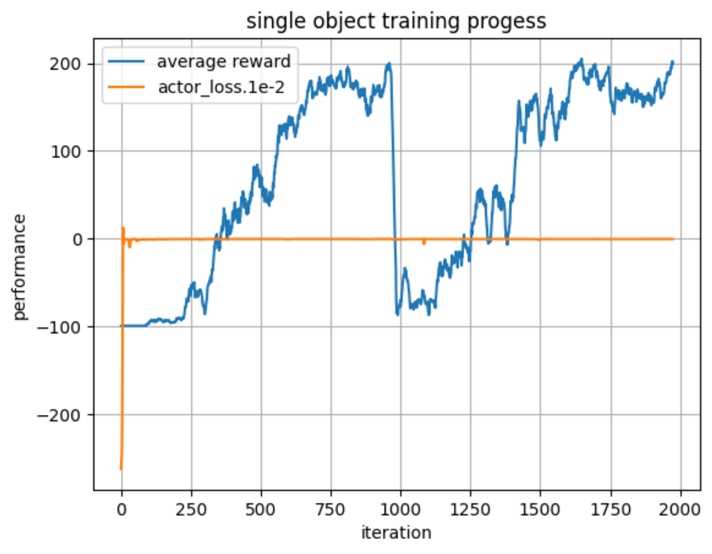
```
Initialize θ to a random network and D ← {}
Initialize R(ob) ← {Rshaping, InverseRL}
while true do
Choose a goal r ~ R(ob)
   for t ∈ {0, ..., T} do
Get state s_t from environment
Select action a_t = π(·|s_t, r, θ)
Get experience {s_t, a_t, r_t, s_{t+1}} and add to β
   end for
end while=0
```

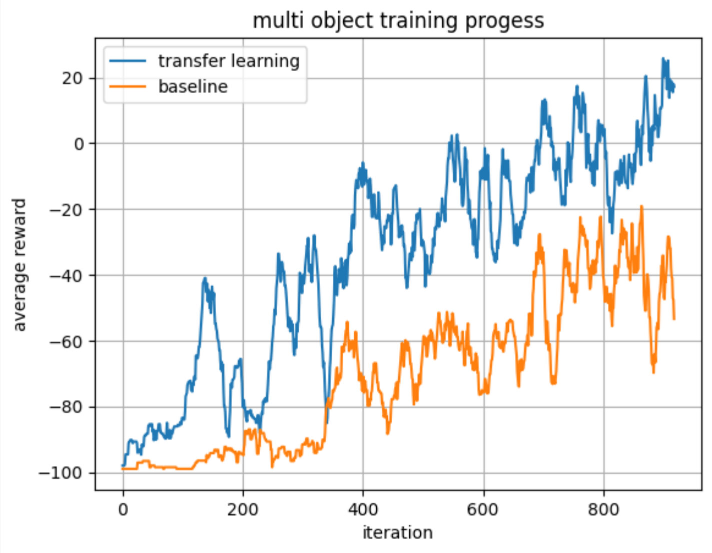
Training process

1. Agent can learn single object pick-and-place easily
2. Multi-object and domain randomization require transfer learning
3. Hierarchical RL approach can form options

Metric	reward	ratio	option_duration
Epoch 823	-67	1.0	[44, 122, 60, 53, 20, 72, 35, 1341]
Epoch 824	-15	2.0	[37, 30, 27, 46, 18, 48, 21, 80]

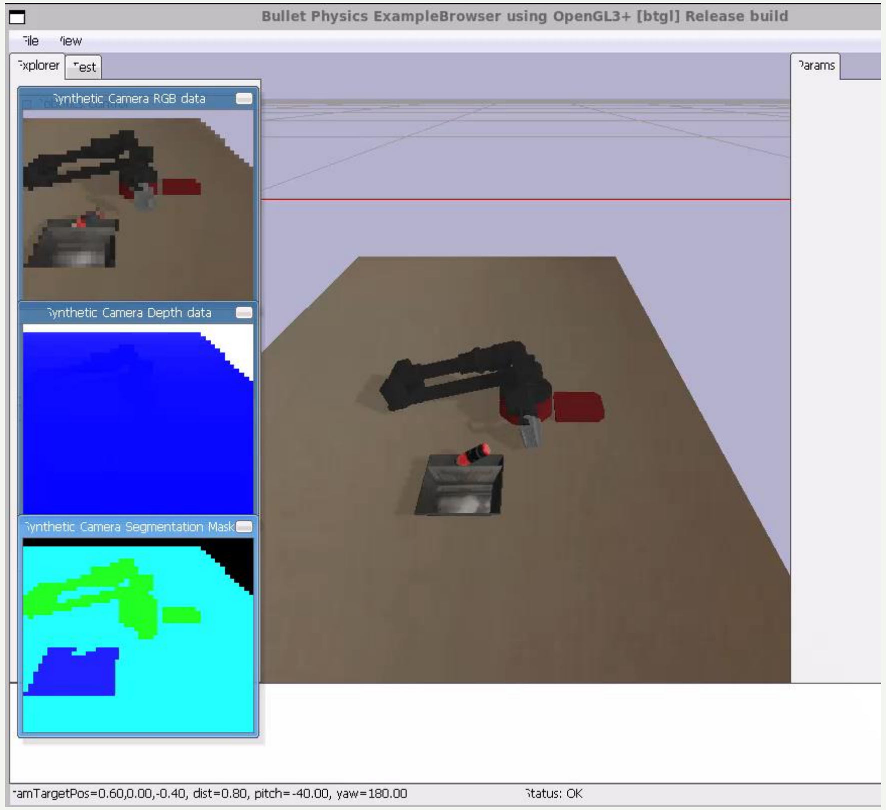
Table 1: Option combination





Result

1. Good single object pick-and-place
2. Mediocre performance in case of multi-object and domain randomization
3. Hierarchical RL online training pipeline



Conclusion

1. Robot object pick-and-place
2. Simple end-to-end goal sequence RL
3. Advanced HER version with reward goal
4. Hierarchical Reinforcement Learning for more complicated problem

Other Discussion items

1. Testing on real robot
   a. Can conduct online inference
   b. Wrong axis setting
2. Hierarchical RL may provide better performance